WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Discrimination of mesophilic and thermophilic proteins using machine learning algorithms

M. Michael Gromiha* and M. Xavier Suresh

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan

ABSTRACT

Discriminating thermophilic proteins from their mesophilic counterparts is a challenging task and it would help to design stable proteins. In this work, we have systematically analyzed the amino acid compositions of 3075 mesophilic and 1609 thermophilic proteins belonging to 9 and 15 families, respectively. We found that the charged residues Lys, Arg, and Glu as well as the hydrophobic residues, Val and Ile have higher occurrence in thermophiles than mesophiles. Further, we have analyzed the performance of different methods, based on Bayes rules, logistic functions, neural networks, support vector machines, decision trees and so forth for discriminating mesophilic and thermophilic proteins. We found that most of the machine learning techniques discriminate these classes of proteins with similar accuracy. The neural network-based method could discriminate the thermophiles from mesophiles at the five-fold cross-validation accuracy of 89% in a dataset of 4684 proteins. Moreover, this method is tested with 325 mesophiles in Xylella fastidosa and 382 thermophiles in Aquifex aeolicus and it could successfully discriminate them with the accuracy of 91%. These accuracy levels are better than other methods in the literature and we suggest that this method could be effectively used to discriminate mesophilic and thermophilic proteins.

Proteins 2008; 70:1274–1279.

Key words: protein stability; machine learning algorithms; neural networks; mesophilic; thermophilic.

INTRODUCTION

Thermophilic organisms produce proteins of extreme stability and they withstand up to the temperature of 120°. The successful discrimination of thermophilic proteins from mesophilic ones is an important problem and it would help to design stable proteins. Several investigations have been carried out to understand the features influencing the stability of thermophilic proteins and are surveyed in detail.[1–6] Based on the comparative analysis of proteins in mesophilic and thermophilic families, Gromiha *et al.*[7] showed that the increase in Gibbs free energy change of hydration ($-G_{hN}$) and shape enhanced the stability of thermophilic proteins. This has been supported by the experimental work of Hasegawa *et al.*[8]; they increased the stability of mesophilic cytochrome c through five substitutions and observed that the $-G_{hN}$ may contribute to the stability. Furthermore, it has been reported that increase in number of salt bridges and side chain–side chain interactions,[9] aromatic clusters,[10] contacts between the residues of hydrogen bond forming capability,[11,12] ion pairs,[13] cation–π interactions,[14,15] noncanonical interactions,[16] electrostatic interactions of charged residues and the dielectric response,[17,18] amino acid coupling patterns,[19] main-chain hydrophobic free energy[20] and hydrophobic residues[21] in thermophilic proteins enhanced the stability.

On the other hand, the amino acid sequences of genomes have been used for understanding the stability of thermophilic proteins. Das and Gerstein[22] analyzed 12 mesophilic and thermophilic families and reported that intrahelical salt bridges are prevalent in thermophiles. Fukuchi and Nishikawa[23] showed that the amino acid composition on protein surface may be an important factor for understanding the stability. Ding *et al.*[24] revealed the preferences of dipeptides in thermophilic proteins for extreme stability. Recently, Berezovsky *et al.*[25] analyzed the amino acid compositions of designed model sequences and natural proteomes and reported that there is a specific trend in the amino acid compositions in response to the requirement of stability at elevated environmental temperature. They revealed that the proteomes of thermophilic proteins are enriched in hydrophobic and charged amino acids at the expense of polar ones.

Zhang and Feng[26] utilized the information about dipeptide composition and developed a statistical method for discriminating mesophilic and thermophilic proteins, similar to the one proposed for discriminating β-barrel membrane proteins.[27] This method showed an accuracy of 86% in dis-

*Correspondence to: M. Michael Gromiha, Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan. E-mail: michael-gromiha@aist.go.jp
Received 9 February 2007; Revised 19 April 2007; Accepted 3 May 2007
Published online 17 September 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21616

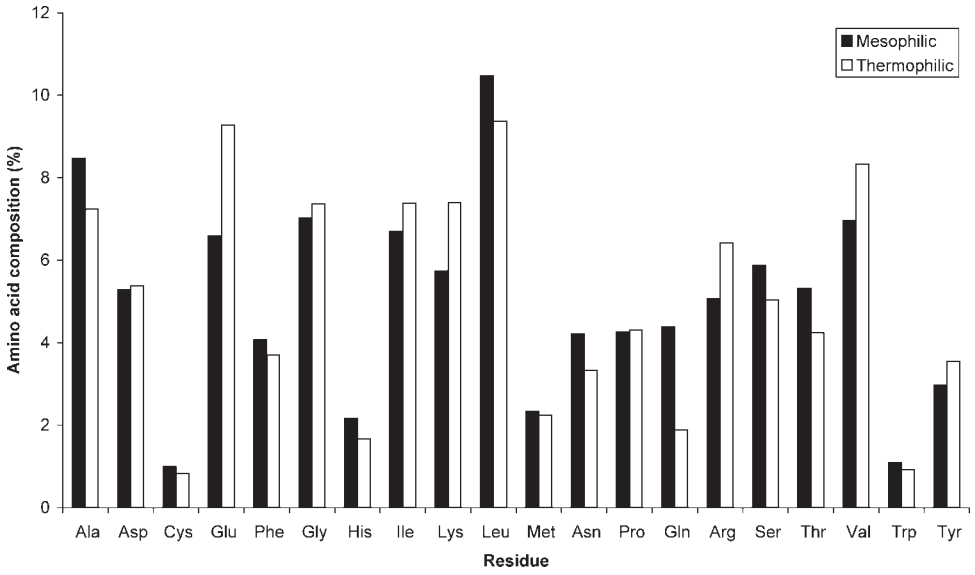

Figure 1

Amino acid composition in mesophilic (■) and thermophilic (□) proteins.

criminating mesophiles and thermophiles. In this work, we have analyzed the performance of different algorithms, such as Bayes rules, neural network, SVM, decision trees and so forth for discriminating mesophilic and thermophilic proteins. We found that the five-fold cross-validation accuracy is almost similar in most of the machine learning algorithms and the accuracy of discriminating mesophilic and thermophilic proteins using neural networks is marginally better than other methods. It could discriminate them at an accuracy of 93 and 89%, respectively, for self-consistency and five-fold cross-validation tests in a dataset of 4684 proteins. Further, the influence of different families will be discussed.

MATERIALS AND METHODS

Datasets

Zhang and Feng[26] used 4895 mesophilic and 3522 thermophilic proteins for discriminating them using dipeptide composition. The proteins in each set contain many redundant sequences and we removed the redundancy using CD-HIT algorithm[28] as implemented by Holm and Sander.[29] The final dataset contains 3075 mesophilic proteins and 1609 thermophilic proteins. Further, we have used a test set of 325 mesophilic and 382 thermophilic proteins belonging to *Xylella fastidosa* and *Aquifex aeolicus* families, respectively. These datasets have the proteins with less than 40% sequence identity.

Computation of amino acid composition

The amino acid composition for each protein has been computed using the number of amino acids of each type and the total number of residues. It is defined as:

$$\mathrm{Comp}(i) = \sum n_i/N \qquad (1)$$

where i stands for the 20 amino acid residues. n_i is the number of residues of each type and N is the total number of residues. The summation is through all the residues in the particular protein. We have repeated the calculation for all the proteins in mesophilic (and thermophilic) organisms and computed the average to estimate the composition of each amino acid residue in mesophilic (and thermophilic) proteins (Fig. 1).

n-fold cross-validation method

We have performed n-fold cross-validation test for assessing the validity of the present work. In this method, the dataset is divided into n groups, $n - 1$ of them are used for training and the rest is used for testing the method. The same procedure is repeated for n times and the average is computed for obtaining the accuracy of the method. We have carried out two-fold, three-fold, four-fold, five-fold, and ten-fold cross validation tests.

Calculation of sensitivity, specificity, and accuracy

We have used different measures to assess the accuracy of discriminating mesophilic and thermophilic proteins. The term, sensitivity shows the correct prediction of thermophiles, specificity about the mesophilies and accuracy indicates the overall assessment. These terms are defined as follows:

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP})$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}),$$

where, TP, FP, TN, and FN refer to the number of true positives (thermophiles identified as thermophiles), false positives (mesophilies identified as thermophiles), true negatives (mesophiles identified as mesophiles), and false negatives (thermophiles identified as mesophiles), respectively.

Machine learning techniques

We have analyzed several machine learning techniques implemented in WEKA program[30] for discriminating mesophilic and thermophilic proteins. This program includes several methods based on Bayes functions, neural networks, logistic functions, support vector machines, regression analysis, nearest neighbor methods, meta learning, decision trees, and rules. The details of these methods have been explained in our earlier article.[31] We have analyzed different classifiers and datasets to discriminate mesophilic and thermophilic proteins.

RESULTS AND DISCUSSION

Statistical analysis of amino acid compositions in mesophilic and thermophilic proteins

We have computed the amino acid composition of mesophilic and thermophilic proteins and the results are shown in Figure 1. From this figure, we observed that the residues Ala, Leu, Gln, Thr, Glu, Lys, Arg, and Val show subtle difference ($|\text{comp}_{\text{thermo}} - \text{comp}_{\text{meso}}| > 1.0$) between mesophiles and thermophiles. Further, the compositional differences of these residues are found to be statistically significant ($P \leq 1.0\text{e}{-6}$). This has been done by comparing the composition of each residue in all the proteins belonging to mesophilic and thermophilic organisms as used in other studies.[32] We have used the program available at http://www.fon.hum.uva.nl/Service/Statistics/2Sample_Student_t_Test.html for calculating the *P*-value. While the composition of Ala, Leu, Gln, and Thr are higher in mesophiles than thermophiles an opposite trend is observed for Glu, Lys, Arg, and Val. These preferences and the higher occurrence of other amino acids in thermophilic proteins reveal the implications for protein stability.

The comparative analysis on the occurrence of Cys, Ile, and Val in the structural homologues of 23 mesophilic and thermophilic proteins[20] showed that the occurrence of Cys is less in thermophiles than mesophiles. On the other hand, the occurrence of Val/Ile is higher in thermophiles than mesophiles. In addition, it has been reported that Cys can be replaced by Val/Ile to enhance the stability.[7] Interestingly, these trends were reflected in the analysis of amino acid composition.

Further, the charged residues, Lys, Arg, and Glu have significantly higher occurrence in thermophilic proteins than mesophilic ones ($|\text{comp}_{\text{thermo}} - \text{comp}_{\text{meso}}| > 1.0$) and the composition of Asp showed a moderate difference (Fig. 1). We have analyzed the composition of charged residues in the structural homologues of thermophilic and mesophilic proteins and observed that the thermophiles have more number of charged residues than mesophiles. This result supports our observation obtained with amino acid sequence analysis.

Discrimination of mesophilic and thermophilic proteins

We have analyzed the performance of different machine learning techniques for discriminating mesophilic and thermophilic proteins. In this discrimination, we have used the amino acid composition as the main attributes. It has been shown that amino acid composition could discriminate DNA and RNA binding proteins, outer membrane proteins, transmembrane helical proteins and so forth with reliable accuracy.[31,33–35] The results obtained for a set of machine learning techniques using five-fold cross validation method are presented in Table I. We observed that most of the machine learning methods discriminated the mesophilic and thermophilic

Table I
Discrimination of Mesophilic and Thermophilic Proteins Using Different Machine Learning Approaches

	Five-fold cross-validation		
	Sensitivity	Specificity	Accuracy
Method	(%)		
Bayesnet	81.4	90.6	87.4
Naive Bayes	83.5	88.8	87.0
Logistic function	82.8	92.8	89.4
Neural network	82.4	93.0	89.4
RBF network	80.7	89.6	86.5
Support vector machines	82.2	92.9	89.2
k-nearest neighbor	77.3	88.7	84.8
Bagging meta learning	80.0	92.0	87.9
Classification via Regression	79.3	91.0	87.0
Decision tree J4.8	75.8	88.4	84.0
NBTree	79.2	89.5	86.0
Partial decision tree	81.5	85.2	83.9

DOI 10.1002/prot

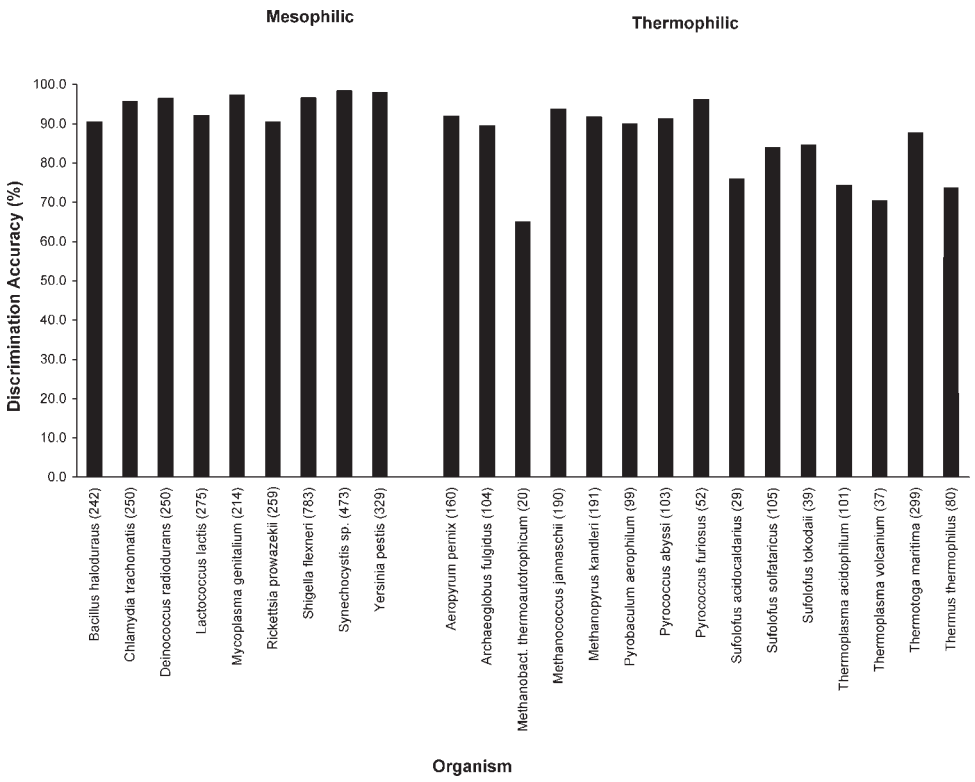

Figure 2

Discrimination accuracy in different mesophilic and thermophilic organisms. The number of proteins in each organism is shown in parenthesis.

proteins with the accuracy in the range of 84–89% in a set of 4684 proteins. This analysis showed that there is no significant difference in performance between different machine learning methods. Interestingly, the methods neural networks, support vector machines and logistic functions discriminated mesophilic and thermophilic proteins at similar accuracy of 89%. Further, the method based on neural networks performed well in discriminating DNA binding proteins, β-barrel membrane proteins, predicting protein secondary structures and so forth.[31,36,37] and hence we used this method for further analysis. The accuracy of identifying thermophilic proteins is 82% where as that of excluding mesophilic proteins is 93%. The overall accuracy is 89.4% for distinguishing mesophilic and thermophilic proteins.

Influence of different families for discrimination

The accuracy of discriminating mesophilic and thermophilic proteins in different families has been analyzed and the results are depicted in Figure 2. In this analysis, we have kept the proteins in a specific organism as a test set and utilized the remaining proteins for training. We observed that the proteins in most of the mesophilic families are discriminated with the accuracy of more than 90%. Interestingly, the proteins from Synechocystis sp. are discriminated with 98.3% accuracy. On the other hand, the accuracy of discriminating thermophilic proteins showed a vide variation of 65–96%. The proteins belonging to *Pyrococcus furiosus* are discriminated with 96% accuracy whereas *Methanobacterium thermoautotrophicum* are discriminated with the accuracy of 65%. Further analysis on this family of proteins revealed that the number of proteins in this family is significantly less (20 proteins) and most of the proteins are showing high sequence identity with mesophilic proteins.

Further, we have analyzed the discrimination accuracy of thermophilic (moderate) and hyper (extreme) thermophilic proteins from mesophilic proteins. Interestingly, we observed that hyper-thermophilic proteins are discriminated with higher accuracy than moderate thermophilic proteins. The accuracies of discriminating hyper-thermophilic and thermophilic proteins from mesophilic ones are, 90 and 73%, respectively.

Performance of *n* -fold cross validation methods

We have analyzed the performance of the present method using different cross-validation methods, ranging

Table II
Influence of Crossvalidation Methods for Discrimination

n-*Fold*	Sensitivity	Specificity (%)	Accuracy
2	79.6	94.4	89.3
3	82.7	91.8	88.7
4	82.5	92.3	88.9
5	81.6	92.9	89.0
10	83.3	92.0	89.0

from 2 to 10 and the results are presented in Table II. We observed a marginal variation in the sensitivity of identifying the thermophilic proteins and the overall accuracy is similar in all the cross-validation methods. It varies in the range of 88.7–89.3%. This result indicates that the mesophilic and thermophilic proteins are discriminated with high confidence and the prediction results are reliable.

Role of amino acid composition and residue pair preference

We have estimated the accuracy of discriminating mesophilic and thermophilic proteins using different attributes, namely, amino acid composition, residue pair preference, and the combination of them. As discussed in the earlier section, the composition of 20 amino acid residues could discriminate the mesophiles and thermophiles at the five-fold cross validation accuracy of 89% (Table I).

It has been reported that the dipeptide composition (residue pair preference) could discriminate the β-barrel membrane proteins and thermophilic proteins at high accuracy.[26,27] Recently, Zhang and Feng[26] analyzed the dipeptide compositions of mesophilic and thermophilic proteins and reported that the dipeptides, EE, KK, RR, PP, KI, VV, VE, KE, VK, QQ, AA, EQ, LL, QA, QL, NN, KQ, QG, RQ, QT, and AQ have significant differences between them ($|dipep_{thermo} - dipep_{meso}| > 1.0$). We have used the compositions of these 21 residue pairs for discriminating mesophilic and thermophilic proteins and obtained the accuracy of 85%. Further, we have tried to combine the compositions of both 20 amino acids and 21 selected dipeptides for discrimination and observed that there is no significant improvement in the accuracy (<1%).

Discrimination of mesophilic and thermophilic proteins with different datasets

We have assessed the reliability of the present method by discriminating mesophilic and thermophilic proteins from different families that are not considered in the work for training/testing. We have collected the data of 325 mesophilic and 382 thermophilic proteins from *Xylella fastidosa* and *Aquifex aeolicus* families, respectively. We observed that the present method based on neural networks correctly identified the thermophilic proteins with the sensitivity of 87.6%. Further, the mesophilic proteins are excluded with the specificity of 95.7% and the overall accuracy is 91.3%. These results demonstrated that our method is performing extremely well in distinguishing mesophilic and thermophilic proteins.

Comparison with other methods

Zhang and Feng[26] proposed a statistical method based on amino acid and dipeptide compositions for discriminating mesophilic and thermophilic proteins and reported the accuracy of 74 and 86%, respectively, using self-consistency test. Our method significantly improved the accuracy up to 93% for discriminating them using amino acid composition. The five-fold cross validation accuracy is 89%, which is better than the accuracy reported with self-consistency test.[26] On the other hand, the number of variables is significantly less (20) in the present method compared with 400 residue pairs used in Zhang and Feng.[26]

The results on the test sequences showed that the dipeptide composition could discriminate them with the accuracy of 89.7% while the present method discriminated the test set of 707 proteins at 91.3% accuracy. These results emphasize that the present method is superior to other methods in the literature.

CONCLUSIONS

We have revealed the amino acid compositional difference between mesophilic and thermophilic proteins. All the charged residues Lys, Arg, Glu, and Asp and the hydrophobic residues Val and Ile have higher occurrence in thermophilic proteins compared with mesophiles. The neural network based method using amino acid composition distinguished the mesophilic and thermophilic proteins at the five-fold cross validation accuracy of 89% for a dataset of 4684 proteins. Further, we could achieve the accuracy of 91% in two independent dataset of 325 mesophiles and 382 thermophiles. We suggest that our method can be used as an effective tool to discriminate mesophilic and thermophilic proteins.

ACKNOWLEDGMENTS

We thank the reviewers for constructive comments on our manuscript.

REFERENCES

1. Ladenstein R, Antranikian G. Proteins from hyperthermophiles: stability and enzymatic catalysis close to the boiling point of water. Adv Biochem Eng Biotechnol 1998;61:37–85.
2. Jaenicke R, Bohm G. The stability of proteins in extreme environments. Curr Opin Struct Biol 1998;8:738–648.
3. Szilagyi A, Zavodszky P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Struct Fold Des 2000;8:493–504.

DOI 10.1002/prot

4. Kumar S, Nussinov R. How do thermophilic proteins deal with heat? Cell Mol Life Sci 2001;58:1216–1233.
5. Yano JK, Poulos TL. New understandings of thermostable and peizostable enzymes. Curr Opin Biotechnol 2003;14:360–365.
6. Razvi A, Scholtz JM. Lessons in stability from thermophilic proteins. Protein Sci 2006;15:1569–1578.
7. Gromiha MM, Oobatake M, Sarai A. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys Chem 1999;82:51–67.
8. Hasegawa J, Uchiyama S, Tanimoto Y, Mizutani M, Kobayashi Y, Sambongi Y, Igarashi Y. Selected mutations in a mesophilic cytochrome c confer the stability of a thermophilic 9. counterpart. J Biol Chem 2000;275:37824–37828.
9. Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. Protein Eng 2000;13:179–191.
10. Kannan N, Vishveshwara S. Aromatic clusters: a determinant of thermal stability of thermophilic proteins. Protein Eng 2000;13:753–761.
11. Gromiha MM. Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. Biophys Chem 2001;91: 71–77.
12. Gromiha MM, Selvaraj S. Inter-residue interactions in protein folding and stability. Prog Biophys Mol Biol 2004;86:235–277.
13. Kumar S, Tsai CJ, Nussinov R. Thermodynamic differences among homologous thermophilic and mesophilic proteins. Biochemistry 2001; 40:14152–14165.
14. Gromiha MM, Thomas S, Santhosh C. Role of cation-pi interactions to the stability of thermophilic proteins. Prep Biochem Biotechnol 2002;32:355–362.
15. Chakravarty S, Varadarajan R. Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. Biochemistry 2002;41:8152–8161.
16. Ibrahim BS, Pattabhi V. Role of weak interactions in thermal stability of proteins. Biochem Biophys Res Commun 2004;325:1082–1089.
17. Xiao L, Honig B. Electrostatic contributions to the stability of hyperthermophilic proteins. J Mol Biol 1999;289:1435–1444.
18. Dominy BN, Minoux H, Brooks CL, III. An electrostatic basis for the stability of thermophilic proteins. Proteins 2004;57:128–141.
19. Liang HK, Huang CM, Ko MT, Hwang JK. Amino acid coupling patterns in thermophilic proteins. Proteins 2005;59:58–63.
20. Saraboji K, Gromiha MM, Ponnuswamy MN. Importance of mainchain hydrophobic free energy to the stability of thermophilic proteins. Int J Biol Macromol 2005;35:211–220.
21. Sadeghi M, Naderi-Manesh H, Zarrabi M, Ranjbar B. Effective factors in thermostability of thermophilic proteins. Biophys Chem 2006;119:256–270.
22. Das R, Gerstein M. The stability of thermophilic proteins: a study based on comprehensive genome comparison. Funct Integr Genomics 2000;1:76–88.
23. Fukuchi S, Nishikawa K. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. J Mol Biol 2001;309:835–843.
24. Ding Y, Cai Y, Zhang G, Xu W. The influence of dipeptide composition on protein thermostability. FEBS Lett 2004;569:284–288.
25. Berezovsky IN, Zeldovich KB, Shakhnovich EI. Positive and negative design in stability and thermal adaptation of natural proteins. PLoS Comput Biol 2007;3:e52.
26. Zhang G, Feng B. Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins. Process Biochem 2006;41:1792–1798.
27. Gromiha MM, Ahmad S, Suwa S. Application of residue distribution along the sequence for discriminating outer membrane proteins. Comput Biol Chem 2005;29:135–142.
28. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 2001;17:282–283.
29. Holm L, Sander C. Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 1998;14:423–429.
30. Witten IH, Frank E. Data mining: practical machine learning tools and techniques, 2nd ed. San Francisco: Morgan Kaufmann; 2005.
31. Gromiha MM, Suwa M. Discrimination of outer membrane proteins using machine learning algorithms. Proteins 2006;63:1031–1037.
32. Saha S, Raghava GPS. Prediction of neurotoxins based on their function and source. In Silico Biol 2007;7:0025.
33. Gromiha MM, Suwa M. A simple statistical method for discriminating outer membrane proteins with better accuracy. Bioinformatics 2005;21:961–968.
34. Bhardwaj N, Langlois RE, Zhao G, Lu H. Kernel-based machine learning protocol for predicting DNA-binding proteins. Nucleic Acids Res 2005;33:6486–6493.
35. Yu X, Cao J, Cai Y, Shi T, Li Y. Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. J Theor Biol 2006;240:175–184.
36. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
37. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition , sequence and structural information. Bioinformatics 2004;20:477–486.